Predicting the Profitability of Out-of-Print Comic Books and Graphic Novels

Greg Muir

DS796

4/27/2020

Abstract

Objective- The goal of this project was to develop a regression model to predict secondary market prices of out-of-print comic books and graphic novels. This would enable the client to identify the most profitable comics to purchase from distributors.

Methods- The datasets utilized were full and low stock datasets originating from Diamond Comic Distributors. Natural language processing techniques were applied to the title fields so they could be fed into the ISBN database API. Jaccard similarity indexes were used to find appropriate matches between the Diamond datasets and the ISBN database. When matches occurred, secondary market prices were generated. Regression models were trained on both the full stock and low stock datasets.

Results- Natural language processing and utilization of Jaccard similarity indexes led to 47.5% and 49.1% matches between secondary market prices generated from the ISBN database and titles from the Diamond low stock and full stock databases. The full stock database was chosen as the most promising dataset based on r^2 scores. A random forest regressor achieved a 10-fold cross validation r^2 score of .258 =/- .019 on the full stock dataset. Predictions were made using the model on the test set. When results were filtered to only included predicted markups greater than 1, the percentage of actual markups greater than 1 rose from .249 to .608. Similarly, the average profit per item rose from -$2.69 to +$3.14.

Conclusions- Utilizing natural language processing techniques and the ISBN database API has proved to be an effective tool for identifying secondary market prices for comic books and graphic novels. Using a random forest regressor and filtering by predicted markup yields large improvements in likelihood of selecting a profitable product and average profitability of products. As the predicted markup threshold increases the percentage of selecting a profitable product increases, but the number of available options is greatly narrowed. This method allows for flexibility in business priorities between maximizing profit on a narrow range of items or stocking a larger subset of profitable items.

Table of Contents

List of Figures

1. Introduction

The client for this project is Steve Hornstein, the owner of mybargaincomics.com. Mybargaincomics is an online retailer of out-of-print comic books and graphic novels. Comic books are considered out-of-print when they are no longer available from a distributor. The distributor that the client uses is Diamond Comic Distributors, the largest distributor in North America. While comics are available from a distributor, they are considered in-print. Because the supply is plentiful, prices tend to stay relatively stable. When comics go out of print, the supply is fixed. The only way to purchase the comic is through the secondary market. The secondary market consists of comic retailers like our client, as well as individuals and stores selling on marketplaces such as Amazon, eBay, and AbeBooks.

The client does not own a brick and mortar location and as such does not receive the foot traffic that brick and mortar locations do. That along with the cost of shipping make dealing in high volume low margin items like in-print comics cost prohibitive for our client. For those reasons, the client prioritizes the higher margins possible with out-of-print comics. Not all comics appreciate in value. Our data will show that the majority of comics available from Diamond sell for less on the secondary market than the wholesale cost from the distributor. It is important for our client to be able to identify which comics will appreciate and which will not so that they can choose which comics to stock.

We will be building a regression model to predict the secondary market price markup of comic books. We will be calculating the ratio of secondary market price to cost available from Diamond and using that markup as our target variable. The resulting model will allow the client to filter available comics by predicted markup and select the highest results. By only selecting the most profitable comics, they should be able to increase profit margins and make a larger impact on the market. Because a regression model gives a continuous output, it will enable the client to filter possible purchases on a sliding scale. If the client wants higher accuracy, they could set a higher threshold of predicted markup to consider. If they want a broader range of options, they could consider a lower markup threshold. A successful model could also be applied to other collectables. The client could expand their business into trading cards, action figures, models, or other similar collectables. With additional data, the model would also allow the client to identify comics available on the secondary market that are undervalued. That would allow them to purchase books from other retailers to be sold at a later date for a profit.

This is especially relevant in the current economic climate. This report is being prepared in May of 2020 when the world is experiencing a pandemic of COVID-19. It has caused many retail businesses deemed non-essential to close temporarily. Brick and mortar comic book shops have been affected as well as the major comic distributors such as Diamond.[1] This presents an opportunity for online business such as the client's to carve out a larger segment of the market.

[1] https://screenrant.com/diamond-stop-new-comic-shipments/

2. Methods

2.1 Data

The data we will be working with comes from Diamond Comic Distributors. Diamond publishes spreadsheets detailing their full inventory and low stock inventory daily. The specific spreadsheets used in this project were the Diamond low inventory spreadsheet from 9/26/2019 hereafter referred to as Low and the full inventory spreadsheet from 11/22/2019 hereafter referred to as All. These dates were chosen to maximize the time between when the spreadsheet was produced and when the analysis took place. The reasoning being that the further a comic is from its release, the more likely it is to experience price variations due to supply shortages. Attempts were made to find older versions of the Diamond full inventory and low stock inventory spreadsheets, but they were not successful.

The Low and All stock inventory spreadsheets share the same eighteen features. The features we have chosen to further explore are the title, category, date of first shipment, suggested retail price, genre, brand code, price from Diamond before discounts, supplier, length, width, height, weight, and case quantity. The features for Diamond item code, net item indicator, UPC number, and off-site indicator were not included as they either provided redundant information, identifying information, or were close to uniform. We chose to include the low stock indicator feature when working with the All dataset but not the Low dataset, as all items included in Low had been flagged as low stock. Since the two datasets were generated on different dates, the Low dataset had significant differences from the subset of All that had been flagged as low stock.

2.2 ISBN Database API

With regards to price, the Diamond spreadsheets only included the suggested retail price and the price from Diamond before discounts. In order to build a model to predict secondary market prices, we needed to supplement this dataset with secondary market pricing information. We identified the ISBN database API as a tool to acquire that secondary market pricing information. The ISBNdb API can perform ISBN lookups given a title string. Then the API can use that ISBN to pull live price data for the book. The price data, when available, corresponds to the lowest price the book is available on the Amazon and AbeBooks platforms for both new and used books.

2.3 Title Matching

In order to find prices, the ISBNdb API requires a 13-digit ISBN. The API can provide that ISBN13 through a title search. The title field in the Diamond database required some preprocessing before it could be used. The title field contained not just the title, but other information such as Diamond product availability codes and codes for special editions, hardcover, adult content, etc. We tokenized each character in the title strings and made a count of each token. We reviewed all tokens with more than 15 occurrences in the title field and after consulting with our client added all high frequency tokens that did not logically belong in title fields to a stop list. We removed those tokens from the title field. Next, we used a regular expression search to identify all instances where the title field included “Vol X of Y”. We removed the trailing “of Y” from each expression so the API could better identify matches.

After cleaning the title fields, we fed them into the ISBNdb API. The API returned json files with the top 40 results for each search listed. In order to pick the correct match, we considered three criteria, the year the comic was published, the publisher/supplier, and the title. We used regular expressions to extract the years from the date of first shipment field in the Diamond spreadsheets and from the date_published and publish_date fields from the json files the ISBNdb returned and formatted them in YYYY format. For the years, we expected the strings to match perfectly. For the titles and publishers, we expected the strings to be similar but not necessarily an exact match. To identify partial matches, we calculated a Jaccard Similarity Index. The Jaccard Similarity index is a method of comparing two sets. For purposes of string comparisons, it is the number of tokens that are in both strings divided by the number of tokens that are in either string. It gives a result between zero and one, with similar strings being closer to one and dissimilar strings being closer to zero. The formula is as follows:[2]

$$J(X, Y) = |X \cap Y| / |X \cup Y|$$

Setting Jaccard index thresholds accounts for slight variations in formatting allows for matches to be made between "Dark Horse Comics" and "Dark Horse" or "Star Trek Wrath of Khan" and "Star Trek the Wrath of Khan" when checking exact string matches would return no match. We determined a perfect match occurred when the publish year matched exactly, the publisher matched at .5 or greater, and the title matched at .5 or greater. If a perfect match did not occur, we considered other criteria for a match. We considered an imperfect match to have occurred if the publisher matched at .5 or greater and the title matched at .65 or greater, if the date matched exactly and the title matched at .65 or greater, if there was no supplier listed but the year published matched exactly, or if the title matched at .75 or greater.

Using the above criteria, we found matches for 47.48% of books listed in Low and 49.11% of books listed in All. Using the ISBN13 numbers obtained in the initial search, we pulled live prices from Amazon and AbeBooks on 4/4/2020. We used the lowest available price from either supplier in either used or new condition. We used this secondary market price to derive target features for our models. This will be discussed further in Section 4.2, Target Variables.

3. Related Work

We were not able to identify any research papers whose work pertains to using regression models predict comic book or even book prices. However, Wang et al. published a paper in EPJ Data Science exploring a model for predicting book sales.[3] They explored feature importance and established that for fiction books, the most important feature is the imprint the publisher is using. Publishers often publish books under different sub brand names all under the same umbrella. Each sub brand will typically publish books that fall within a specific category like business or children's books. Those sub brands are referred to as imprints. While comic publishers do not use imprints, we expect the publisher to be a close comparison. The second most important feature Wang et al. discovered for fiction books was an author's previous sales. They pointed to evidence that readers tend to choose books by celebrities or by

[2] https://www.statisticshowto.com/jaccard-index/

[3] https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-019-0208-6

authors whose books they have previously read. We expect that with regards to comic books where authors and illustrators change frequently, that this demand for familiarity will play out in readers seeking out other comics that feature the same characters. If Superman comics are popular, a new Superman comic is likely to be popular as well. While Wang et al. were predicting book sales rather than book prices, a high demand is an important component of high secondary market prices.

There are many research papers that deal with using regression models to predict prices, just not book prices. There are many that seek to predict future prices of stocks and goods, but there is a subset that attempt to identify the correct price for an item based on similar items. A paper published by Kalehbasti, Nikolenko, and Rezaei explored several models with the goal of predicting prices of Airbnb listings.[4] They used manual selection, lasso regularization, and lowest p-values during regression to select features, finding lasso regularization to have the best r^2 score. For modeling they tried ridge regression, k-means clustering with ridge regression, support vector regression, neural networks, and gradient boosting tree ensembles and found the support vector regression to yield the best r^2 score and mean squared error. The neural network, ridge regression, and k-means clustering with ridge regression also performed well. Since we are also using regression for price prediction, we intend to explore those models to see which works best for our data.

4. Data Analysis

4.1 Hypothesis

We expect to identify a small signal from the available features driving a markup in prices. We expect that much of the variation in markup price will not be captured by the features in the dataset. We anticipate there to be several outside factors which influence the success of specific comics that are not contained within our data. Factors such as whether a comic is experiencing increased popularity due to a Netflix show/blockbuster movie, whether the art on the cover is especially well done, or whether this issue introduces a character that will go on to be popular are all factors that are thought to influence the price of a comic book[5] that wouldn't be contained in this data. We do expect some of the factors that influence the markup in prices of comic books to be contained within our data. We expect that comics that are volume one in a series will drive a higher price. Additionally, we expect that certain superheroes and suppliers will be popular while others will not. Our client believes that larger special editions tend to be popular and as such we expect the volume to influence markup.

We predict that the Low dataset will show a stronger signal than the All dataset. We attribute this to a belief that items flagged as low stock better resemble the out-of-print comics whose limited supply allows for high secondary market prices. However, we believe that the closure of Diamond Comic Distributors due to the COVID-19 pandemic has placed a limit on supply of all items regardless of Diamond stock. For that reason, we will be testing models on both the Low and All datasets and will use whichever model produces a better r^2 score.

[4] arXiv:1907.12665 [cs.LG]

[5] https://www.comicbookdaily.com/collecting-community/market-trends/factors-that-drive-demand-for-comics/

4.2 Target Variables

We began by deriving target variables from the secondary market data obtained from the ISBNdb API as well as the Diamond All and Low datasets. We calculated a markup feature based on the cost available to our client from Diamond vs the lowest price available online as identified by ISBNdb. Our client receives a 35% discount on products from DC Comics or Marvel Comics, no discount on products from Dark Horse Comics, and a 50% discount on products from all other suppliers. The formula for our markup feature was as follows:

markup = (secondary market price)/ ((base cost from Diamond)*(discount rate))

We looked at histograms of markup from All and Low and determined that markup had a right skew. Natural log transformations were made to make the distribution of the target variable more normal. We used the resulting log_markup feature as our target feature. The histograms are displayed in figures 4.2.1-4.2.4. We removed all observations that fell outside 4 standard deviations from the mean as outliers.

Fig 4.2.1 Distribution of markup in Low

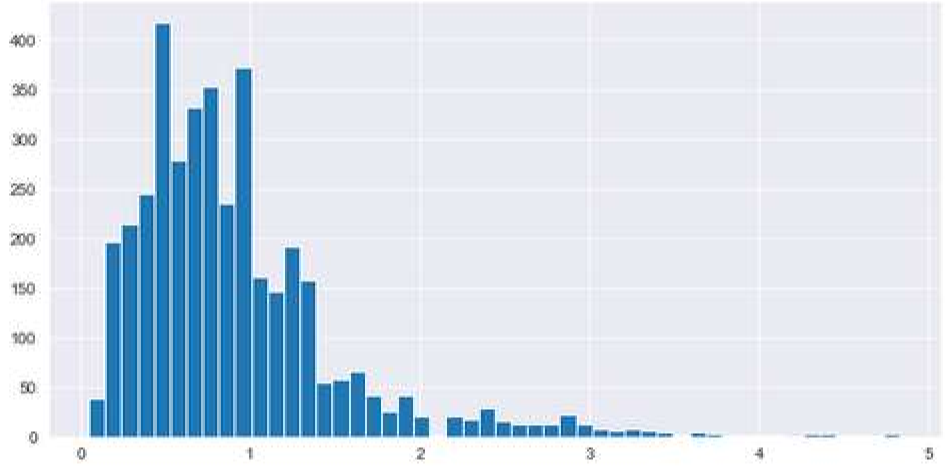


Fig 4.2.2 Distribution of markup in All

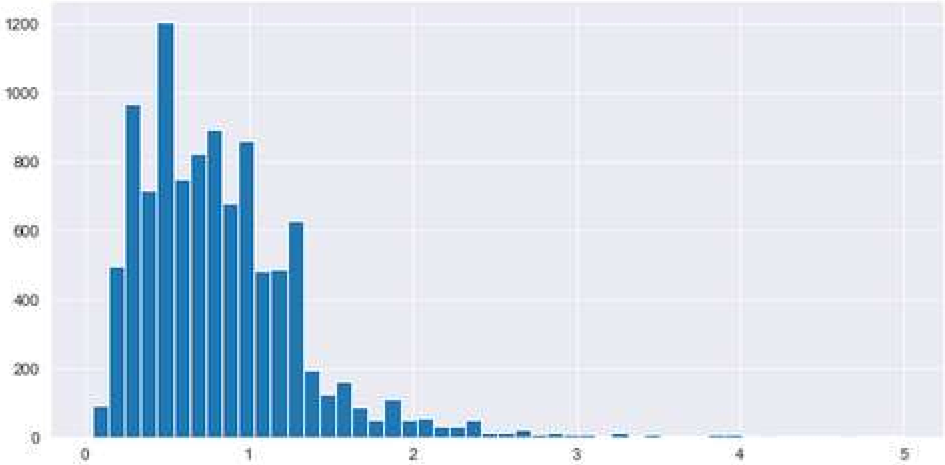


Fig 4.2.3 Distribution of log_markup in Low

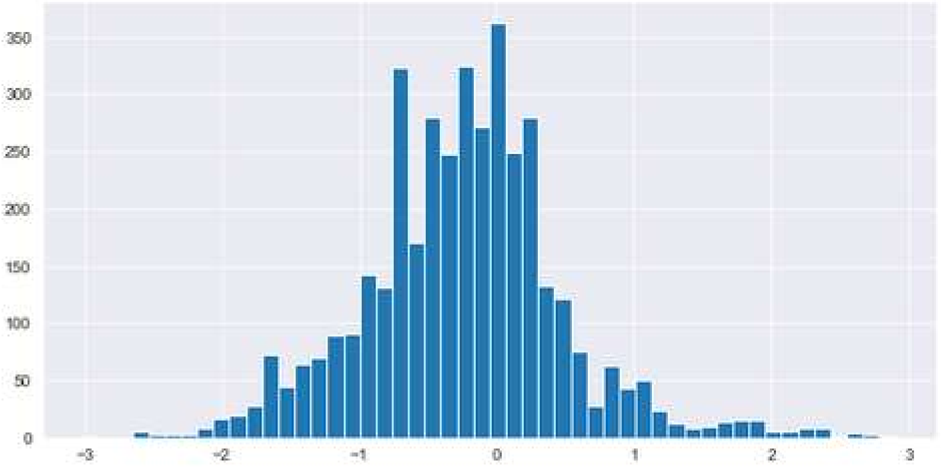
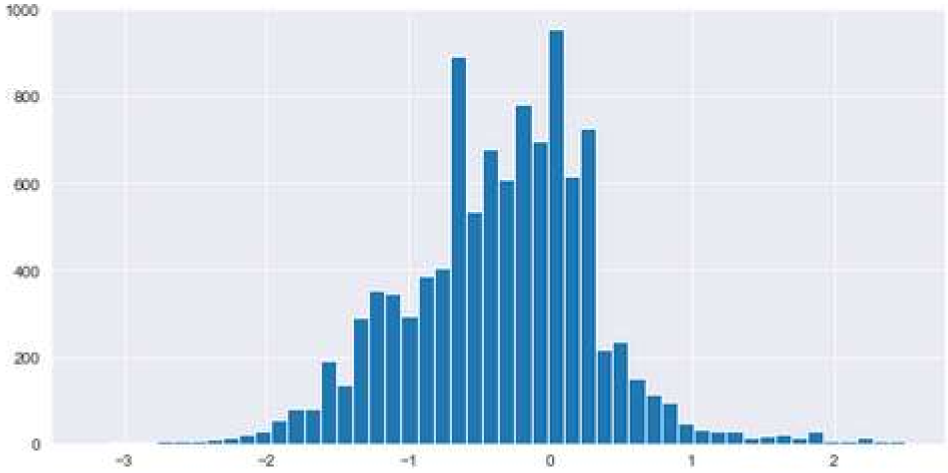


Fig 4.2.4 Distribution of log_markup in All



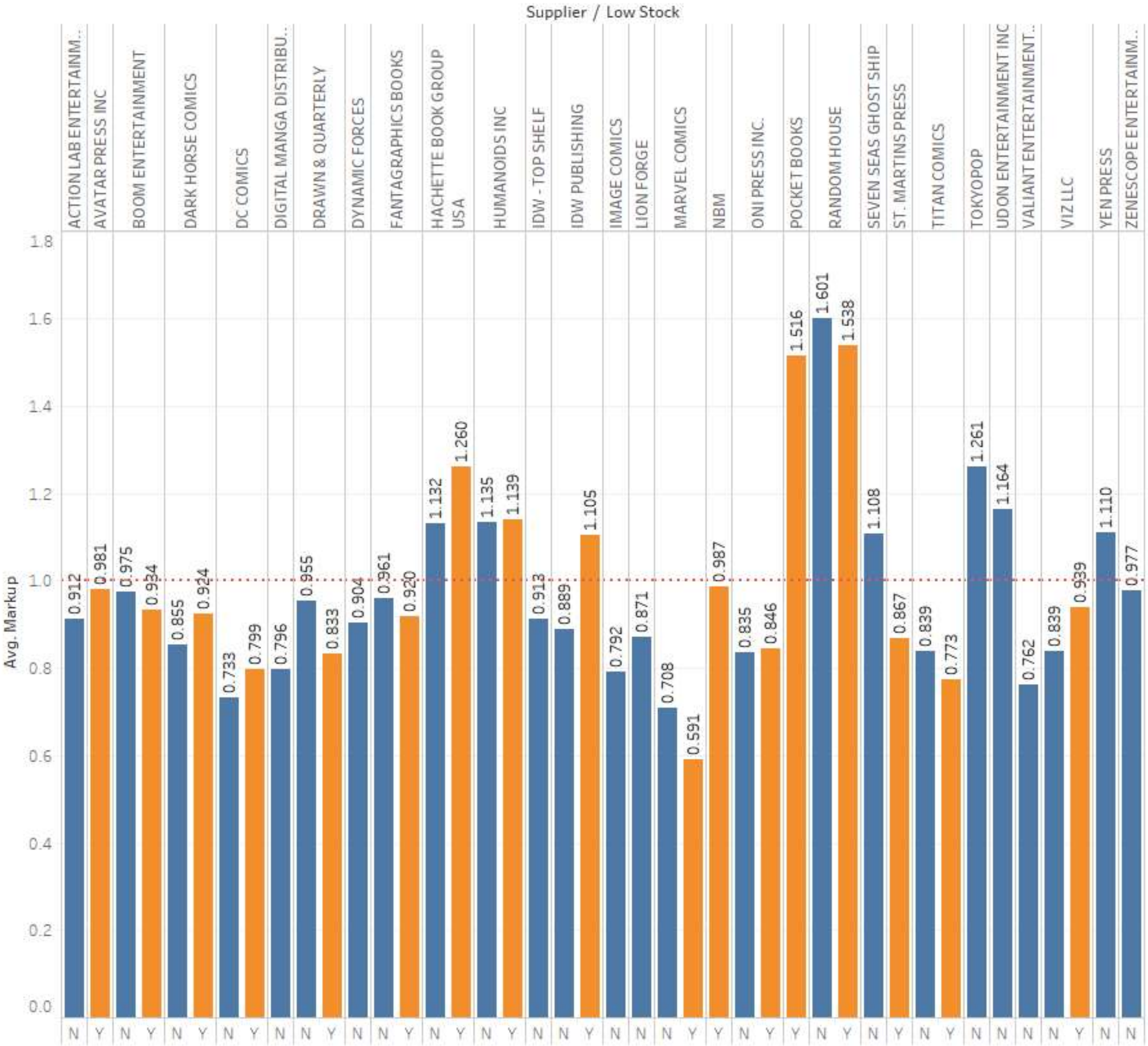
4.3 Feature Engineering

We began feature selection by deriving several features from the datasets. We performed a regular expression search of the title field. Any observation whose title included ‘VOL 1’ was flagged as a 1 in vol_1 all others were assigned a 0. The height, width, and length features were combined into a single volume feature.

Next, we looked at the categorical features. The supplier feature indicates the publisher of the comic. Since our client receives different discount rates based on the supplier, we knew this would be an important feature to include. However, there are 121 unique suppliers in the dataset. Including all of them would likely lead to overfitting. To select the supplier, we eliminated all suppliers who did not have either 25 or more items flagged as low stock, or 25 or more items not flagged as low stock. We then plotted average markup by supplier. The chart is shown in Figure 4.3.1. Using the average markup and number of records, we selected the following 11 suppliers to include: Marvel Comics, DC Comics, Image

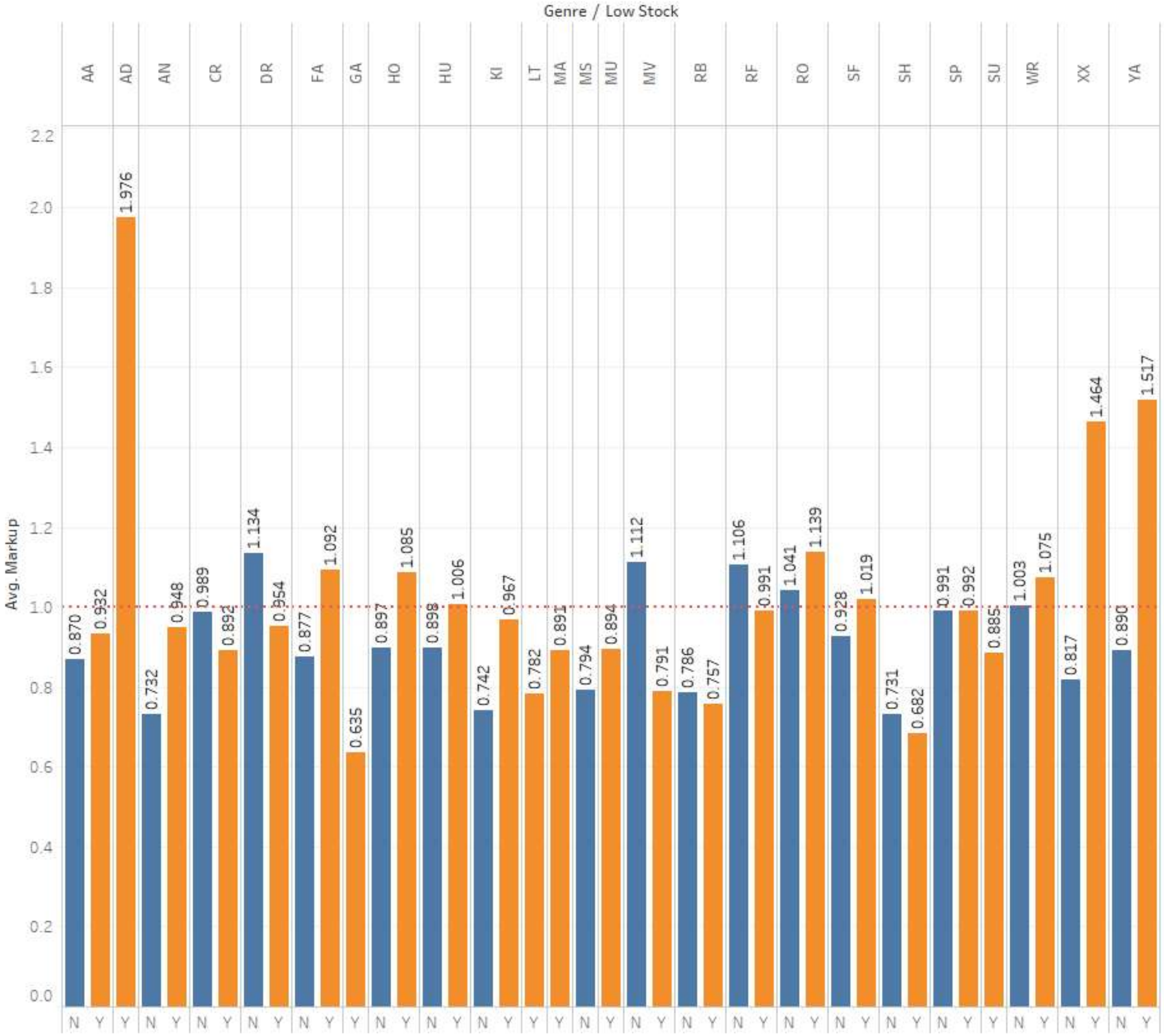
Comics, Dark Horse Comics, Viz LLC, IDW Publishing, Drawn & Quarterly, Hachette Book Group, Oni Press Inc., Pocket Books, and Random House. Features were derived for each, with a 1 assigned if the publisher matched and a 0 assigned otherwise.

4.3.1 Average Markup by Supplier (Low Stock in Orange, Non-Low Stock in Blue)



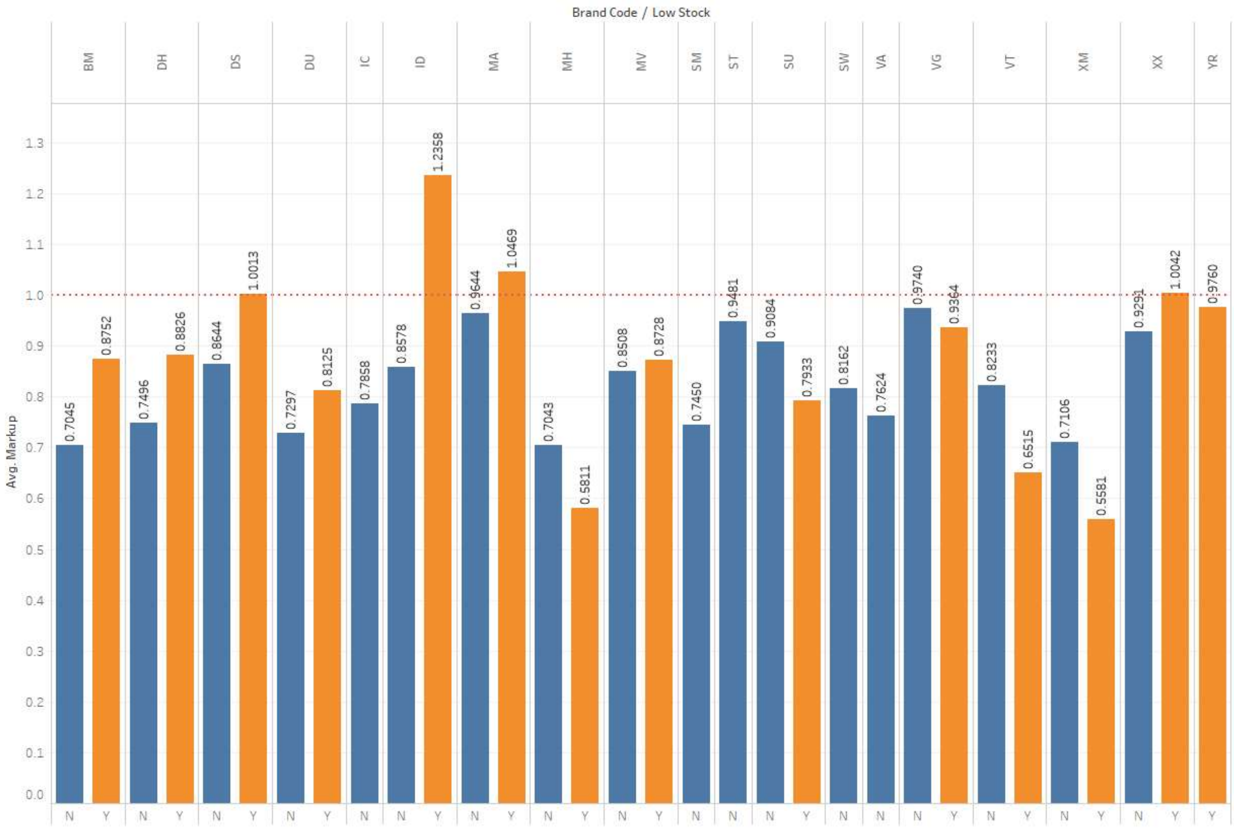
The genre feature is another categorical feature assigned to each item by Diamond. It contains 45 different two letter abbreviations for genres. The key is located in Appendix A. We used the same constraint of 25 or more records in low stock or 25 or more records in non-low stock. We plotted average markup by genre. The chart is shown in Figure 4.3.2. With the same consideration for average markup and number of records, we chose the following 12 genres to include: Superhero, Fantasy, Gaming, Romance, Reality, Horror, Yaoi, Movie, Adult, Literary, Crime, and No Genre. Features were derived for each, with a 1 assigned if the genre matched and a 0 assigned otherwise.

4.3.2 Average Markup by Genre (Low Stock in Orange, Non-Low Stock in Blue)



The brand_code feature is a categorical feature assinged by Diamond to capture popular characters or themes. Each item in the database is assigned a two letter brand_code. The key for brand_code is found in Appendix B. We used the same process as for supplier and genre, eliminating based on a 25 record threshold then plotting average markup by brand_code. The chart is shown in Figure 4.3.3. When examining the chart we noticed potential issues with multicolinearity between the brand_code and supplier/genre. For example, two of the lower performing brand_codes are BM and XM representing Batman and the X-men respectively. Those brands both fall under the Superhero genre and are supplied exclusively by DC Comics and Marvel Comics respectively. Marvel Comics, DC Comics, and the Superhero genre are all underperformers in their respective categories. In order to reduce issues with multicolinearity we have chosen to exclude all brand_codes from our models.

4.3.3 Average Markup by Brand_Code (Low Stock in Orange, Non-Low Stock in Blue)



4.4 Feature Selection

For feature selection, we took an approach similar to Kalehbasti, Nikolenko, and Rezaei[6]. We used three subsets of features. We used a full set of features with identifying features, features with obvious multicollinearity, and features with little discriminatory power removed. From that full set of features, a lasso regression was performed. The features selected by the lasso model were used as a second set of features. The third set of features were selected by hand. The highest r^2 scores resulted from the hand selected features.

4.5 Modeling

A .67-.33 train test split was performed on each dataset before model testing. For regression modeling, we tested 5 models. The regressors we tested were lasso, ridge, support vector, random forest, and k-neighbors. We supplemented those with a Tree-Based Pipeline Optimization Tool (TPOT) that performs model evaluation and hyperparameter tuning[7]. Ultimately, we found the best model based on r^2 score to be a random forest regressor with bootstrap=False, max_feature=.15, min_samples_leaf=2, min_samples_split=17, and n_estimators=100. The model achieved an r^2 score of .258 +/- .019 on 10-

[6] arXiv:1907.12665 [cs.LG]

[7] Trang T. Le, Weixuan Fu and Jason H. Moore (2020). Scaling tree-based automated machine learning to biomedical big data with a feature set selector. *Bioinformatics*.36(1): 250-256.

fold cross validation when trained on the All dataset. The same model achieved an r^2 score of .242 +/- .043 on 10-fold cross validation when trained on the Low dataset. Because the All dataset provided a higher mean r^2 score on validation, we chose to move forward with that model.

4.6 Model Evaluation

We checked diagnostic plots to see how the random forest regressor performed. We plotted observed vs predicted values, residuals vs predicted values, square root(absolute value(standardized residuals))) vs fitted values, and a normal QQ plot. Those charts are shown in Figure 4.6.1. The observed vs predicted plot shows a random distribution about a diagonal line as expected. The residuals vs predicted values plot shows a random distribution about a straight line which is good. The scale location plot shows a slight downward trend but is mostly a random distribution. The normal QQ plot indicates normality with some variation in the top corner. The mean of residuals was calculated to be -.0052 which is sufficiently close to zero. The variance inflation factors of the features used are shown in Figure 4.6.2. The year and Hero flag show vif's of 6.42 and 5.86 respectively. While these are somewhat concerning, they are still below 10 and should still be candidates for inclusion.

Figure 4.6.1 Regression Diagnostic Plots

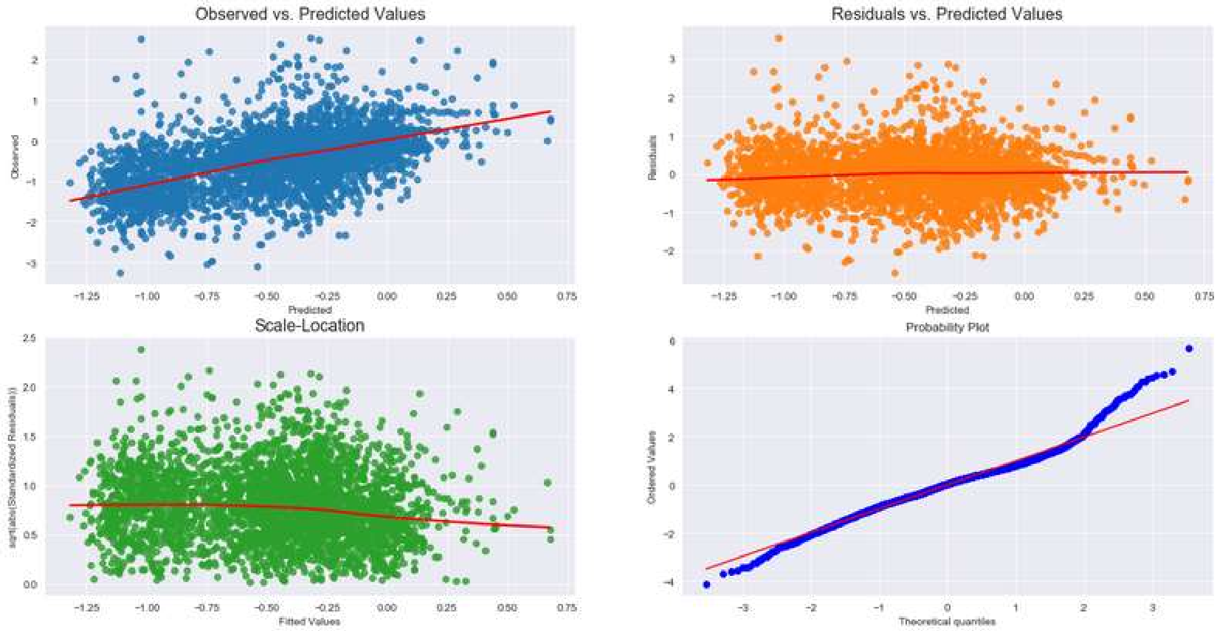
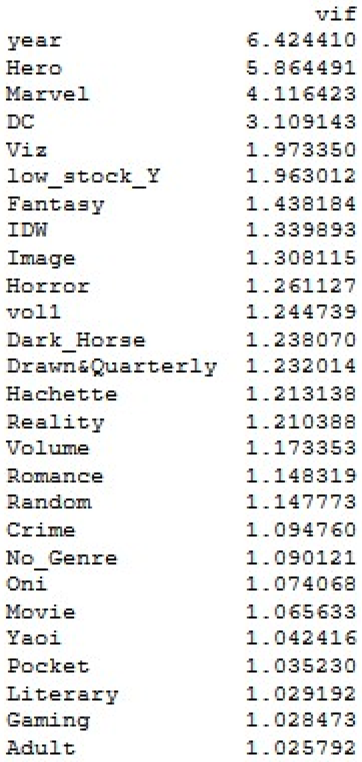


Figure 4.6.2 Variance Inflation Factors of Features



```
                       vif
year              6.424410
Hero              5.864491
Marvel            4.116423
DC                3.109143
Viz               1.973350
low_stock_Y       1.963012
Fantasy           1.438184
IDW               1.339893
Image             1.308115
Horror            1.261127
vol1              1.244739
Dark_Horse        1.238070
Drawn&Quarterly   1.232014
Hachette          1.213138
Reality           1.210388
Volume            1.173353
Romance           1.148319
Random            1.147773
Crime             1.094760
No_Genre          1.090121
Oni               1.074068
Movie             1.065633
Yaoi              1.042416
Pocket            1.035230
Literary          1.029192
Gaming            1.028473
Adult             1.025792
```

5. Assumptions and Risks

A major assumption of this study is that our data accurately reflects typical secondary market prices. This can be broken down into four components. Is this time period an accurate reflection of a typical secondary market economy? Are the comic books listed in the Diamond available spreadsheets a good representation of out of print comics? Are the listed books correctly matched to online prices? Are those prices an accurate reflection of the prices at which people are purchasing those books?

The first major assumption is that this time period is an accurate reflection of a typical secondary market economy. This report is being prepared in April of 2020 when the world is in the midst of a major COVID-19 pandemic. In the United States, at the time of prices were calculated, 42 of 50 states were under shelter in place orders[8]. Comic book stores had been closed as non-essential businesses and 6.6 million Americans filed for unemployment in the week ending March 28th.[9] Diamond Comic Distributors had

[8] https://www.nytimes.com/interactive/2020/us/coronavirus-stay-at-home-order.html

[9] https://www.cnn.com/2020/04/02/economy/unemployment-benefits-coronavirus/index.html

been closed as a non-essential business[10]. While it is not yet evident what the extent of the financial impact of the COVID-19 pandemic will be, in previous times of recession the collectibles market has been negatively affected[11]. We do expect that the comic book secondary market is in a period of lower demand. However, since many comics are normally purchased in brick and mortar stores which are currently closed, we expect there to be an increased demand for comics that are available through online retailers. We are operating under the assumption that the increased demand for comics available from online retailers will counteract the overall lower demand for collectibles during times of recession. Further, we expect that the reduced supply of in print comics from brick and mortar stores and distributors will lead to higher than usual markups for in-print comics. Therefore, we expect to see a stronger signal in the full inventory spreadsheet than we otherwise would. However, we would recommend recalculating prices once the pandemic has subsided and retraining the model based on those prices. That should give a more accurate representation of secondary market prices.

Next, we assume that the prices we have used to calculate the markup for each comic are an accurate representation of the price that customers are paying for those comics. This assumption has two prongs. First, are the prices generated by the ISBNdb API for the correct comic? Finding a price requires two searches through the ISBNdb API. The first search takes a title string and returns 40 possible matches from which a 13-digit ISBN can be extracted. The second search takes one ISBN13 and returns a list of prices. If the wrong ISBN13 is returned from the 40 available options, the resulting price will not accurately reflect the price of the listed title. There are approximately 21,000 comics in the full inventory spreadsheet and 5,000 comics in the low inventory spreadsheet that meet our criteria for inclusion. It would not be feasible to check for correct matches by hand, so a programmatic matching algorithm must be implemented. The details for that can be found on page 6 in the approach to research section. By implementing constraints on title, year of publication, and publisher matches, we have eliminated what we believe to be improper matches and found ISBN13s for approximately 47% of books listed in the low inventory spreadsheet and 49% of books listed in the full inventory spreadsheet. We consider the successful match of title to ISBN to be a true positive. We define a false positive to be a case where we returned an ISBN but for the wrong title and a false negative to be a case where no ISBN was returned even though the correct match was listed. While we expect some false positives to have slipped though, we have worked to find an appropriate tradeoff that also suppresses the false negative rate enough to yield sufficient data. We expect that some of the outliers in markup price are due to these mismatches of title to ISBN. The second prong of this assumption is that the prices returned reflect the price at which customers are purchasing books. The ISBNdb API returns the lowest prices for new and used comics from AbeBooks and Amazon, two of the major online book marketplaces. While those are the lowest listed prices, we have no verification that any books are actually sold at those prices. Since these platforms act as marketplaces for third party merchants to sell items, many items have only a single copy available from a single merchant at the lowest price point. When that item sells, the lowest listed price changes to the next best price. So, the lowest listed price is best thought of as the lowest listed price for which an item has not been sold. For the purposes of this experiment, we are operating under the assumption that the difference between the price comics are selling and the lowest listed price on both platforms is negligible.

[10] https://retailerservices.diamondcomics.com/DiamondDaily/1390/242134

[11] https://www.stltoday.com/business/local/collectibles-market-has-been-hit-hard/article_b474d0b6-9756-5e08-a7be-e803a79361c4.html

6. Future Work

The highest priority for future work is to recalculate secondary market prices at a later date. This will have three benefits to the model. It will increase the time between printing and when the prices were calculated. This will lead to a higher percentage of out-of-print books in the Low and All datasets, which should result in a stronger signal. The second benefit is that it should give a more accurate picture of the current economic climate. While secondary market prices generated during the COVID-10 quarantine reflect the reality of the quarantine, they should not be relied upon to accurately predict prices when there is not a quarantine in place. Third, it allows for a larger training dataset. The initial training datasets can be supplemented with newer ones as time goes on increasing the total number of unique observations that can be used for training.

We anticipate the model could be improved through the inclusion of additional datasets. The data available from Diamond contains no measure of the quality or critical acclaim of the comics. If comic book critic review scores similar to those available on https://comicbookroundup.com/[12] could be added to the dataset it would likely improve the predictive power of the model. Unfortunately, at the time of writing no API existed to interact with the comic book roundup database without web scraping. One other possible inclusion would be to use a method similar to Wang et al.[13] and use Wikipedia page views of authors as a measure of popularity. The ISBNdb includes author information on some of its entries. We chose not to include this method as it would restrict our training data to only the subset of observations where authors could be successfully identified. In the future, once a larger amount of training data is available, this approach could yield improved model accuracy.

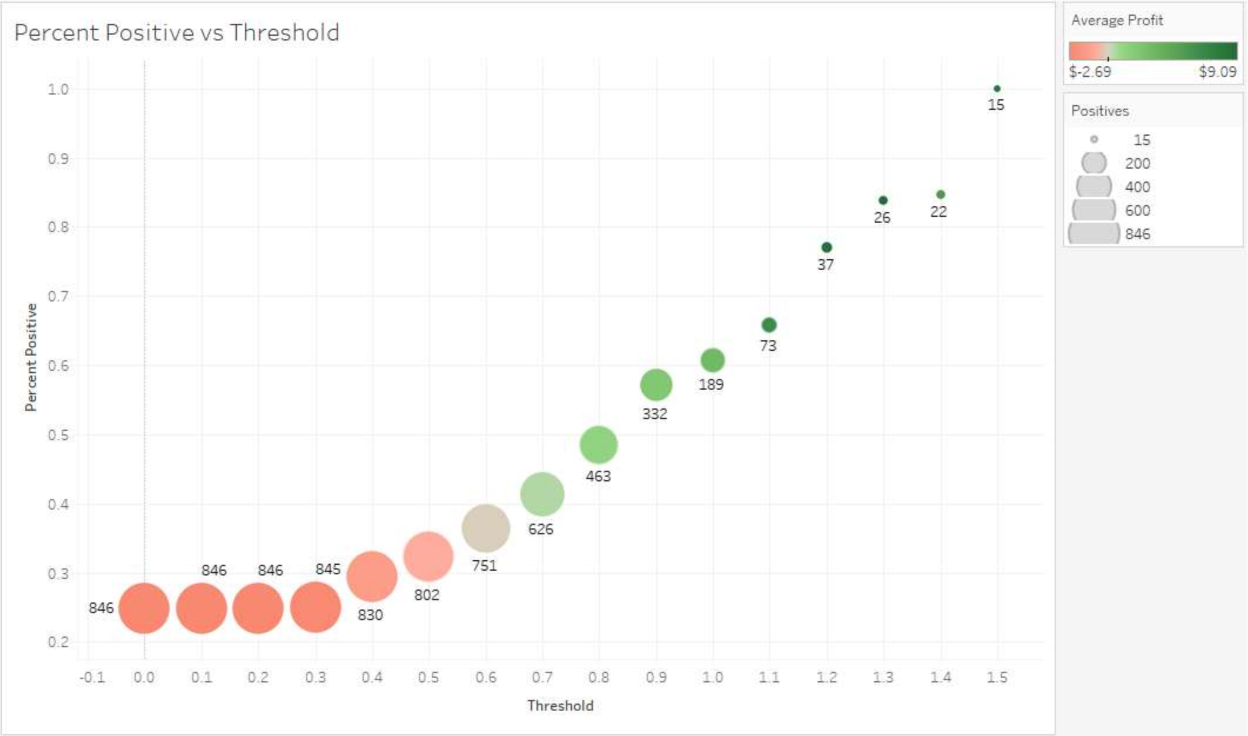
7. Conclusion

The regression model we built achieved a 10-fold cross validation r^2 score of .258 +/- .019 when predicting log_markup. That is more than adequate for our client to increase the profitability of the comics he purchases. In our test set, there were 3404 comics listed, of which 846 had a markup greater than 1. The ratio of markups greater than 1 to those less than or equal to 1 was .249. The average profit in the test set was -$2.69 per item. We applied our regression model to the test set and eliminated all observations with predicted markup of 1 or less. The result was a total of 311 comics with 189 having markups greater than 1 for a ratio of .608. The average profit in the test set became +$3.15 per item. While this method greatly reduces the number of choices our client has when selecting products, it more than doubles the likelihood of turning a profit. The continuous output of the regression model allows the threshold for predicted markup to be easily changed. As the threshold for predicted markup goes up, the percentage of positive observations goes up and the number of positive observations goes down. This interaction can be seen in Figure 7.1. This allows the client to select a threshold suitable for the number of comics they would like to purchase. If the client only wants to purchase 50 unique items to sell, a markup threshold of 1.1 will yield sufficient results and a higher percentage of markups greater than 1 than compared to a threshold of 1.0.

[12] https://comicbookroundup.com/

[13] https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-019-0208-6

Figure 7.1 Percent Positive vs Threshold



8. References

Firestone, A. (2020, March 23). Comic Book Industry's Biggest Distributor To Stop Shipments. Retrieved April 27, 2020, from https://screenrant.com/diamond-stop-new-comic-shipments/

Stephanie. (2019, August 26). Jaccard Index / Similarity Coefficient. Retrieved April 4, 2020, from https://www.statisticshowto.com/jaccard-index/

Wang, X., Yucesoy, B., Varol, O., Eliassi-Rad, T., & Barabási, A.-L. (2019). Success in books: predicting book sales before publication. *EPJ Data Science*, *8*(1). doi: 10.1140/epjds/s13688-019-0208-6

Kalehbasti, P., Nikolenko, L., Rezaeri, H., (2019). Airbnb price prediction using machine learning and sentiment analysis. arXiv: 1907.12665

Factors that Drive Demand for Comics • Comic Book Daily. (2013, April 11). Retrieved April 4, 2020, from https://www.comicbookdaily.com/collecting-community/market-trends/factors-that-drive-demand-for-comics/

Trang T. Le, Weixuan Fu and Jason H. Moore (2020). Scaling tree-based automated machine learning to biomedical big data with a feature set selector. *Bioinformatics*.36(1): 250-256.

Mervosh, S., Lu, D., & Swales, V. (2020, March 24). See Which States and Cities Have Told Residents to Stay at Home. Retrieved April 4, 2020, from https://www.nytimes.com/interactive/2020/us/coronavirus-stay-at-home-order.html

Tappe, A. (2020, April 2). A 3,000% jump in jobless claims has devastated the US job market. Retrieved April 4, 2020, from https://www.cnn.com/2020/04/02/economy/unemployment-benefits-coronavirus/index.html

Diamond Coronavirus Coverage Roundup. (n.d.). Retrieved April 26, 2020, from https://retailerservices.diamondcomics.com/DiamondDaily/1390/242134

Leckey, A. (2011, November 6). Collectibles market has been hit hard. Retrieved April 4, 2020, from https://www.stltoday.com/business/local/collectibles-market-has-been-hit-hard/article_b474d0b6-9756-5e08-a7be-e803a79361c4.html

New Comics - Compare What The Critics Say. (n.d.). Retrieved April 27, 2020, from https://comicbookroundup.com/

Appendix A: Diamond Genre Codes

Genre Code	Description	Genre Code	Description	Genre Code	Description
AA	Action/Adventure	HS	Historical	RF	Reference/Art Books/How To
AD	Adult	HT	How to Draw	RL	Religious
AN	Anthology	HU	Humor/Comedy	RO	Romance
AP	Anthropomorphics	KI	Kids	SF	Science Fiction
AS	Art Supplies	LG	Legend	SH	Super-hero
CJ	Comics Journalism	LT	Literary	SN	Seasonal
CR	Crime	MA	Manga	SP	Sports
DR	Drama	MS	Mystery	SU	Surreal/Non-Linear
DT	Designer Toys	MU	Music	TY	Toy Tie-In
FA	Fantasy	MV	Movie/TV Tie-In	WR	War
GA	Gaming/Role Playing	PC	Pop Culture	WS	Western
HA	Halloween	PK	Pokémon	XX	*No Genre* or *All Genre*
HO	Horror	RB	Reality-Based	YA	Yaoi

Appendix B: Diamond Brand Codes

Brand Code	Brand	Brand Code	Brand	Brand Code	Brand	Brand Code	Brand
AL	Aliens	**DW**	Doctor Who	**MH**	Marvel Heroes	**SW**	Star Wars
AM	Aftermath	**FB**	Football	**MK**	Maverick	**TC**	Top Cow
AN	Anime	**FM**	Frank Miller	**MM**	Mike Mignola	**TM**	Spawn (Todd McFarlane)
BB	Baseball	**GJ**	G.I. Joe	**MT**	Marvel Tech	**UD**	Udon Studios
BH	Bobbing Heads	**GZ**	Godzilla	**MV**	Movie/TV	**VA**	Valiant Heroes
BK	Basketball	**HB**	Highbrow	**NK**	Nickelodeon	**VG**	Video Game/Software Tie-In
BL	Babylon 5	**HC**	Hercules	**PD**	Paradox	**VT**	Vertigo
BM	Batman	**HE**	Helix	**PR**	Predator	**WB**	Web Comic
BU	Buffy the Vampire Slayer	**HK**	Hockey	**PW**	Professional Wrestling	**WK**	WizKids
CC	Candy & Confections	**IC**	Image Comics	**RC**	Racing	**WS**	WildStorm
CI	Classics Illustrated	**ID**	IDW Publishing	**RP**	Roleplaying/D20 Compatible	**WW**	World of Warcraft
CN	Cartoon Network	**IJ**	Indiana Jones	**SF**	Street Fighter	**XE**	Xena
CO	Conan	**JD**	Judge Dredd	**SI**	Studio Ice	**XF**	X-Files
DH	Dark Horse Heroes	**LS**	Lost In Space	**SM**	Spider-Man	**XM**	X-Men
DM	Dark Horse Merchandise	**LT**	Looney Tunes	**SS**	Simpsons	**XX**	*none*
DS	Disney	**MA**	Manga	**ST**	Star Trek	**YA**	Yaoi
DU	DC Universe	**MG**	Magic: The Gathering	**SU**	Superman	**YR**	For Young Readers